Inference for Quantitative Data

Shannon Pileggi

STAT 217

OUTLINE

Beat the Blues

- enrolled patients with depression/anxiety
- randomly assigned them to Treatment as Usual (TAU) or BtheB, a new treatment delivery therapy via computers
- measured depression via Beck Depression Inventory (BDI) at baseline (pre-treatment), and 2, 4, 6, and 8 month follow up
- BDI scores range from 0 to 63 with higher scores indicating more severe depression

Is this study observational or experimental?

1. observational
2. experimental

1. Describe the sample:

2. Describe the population:

The first 6 observations in the data set

```
> head(BtheB)
  drug length treatment bdi.pre bdi.2m bdi.4m bdi.6m bdi.8m
1   No    >6m       TAU      29      2      2     NA     NA
2  Yes    >6m     BtheB      32     16     24     17     20
3  Yes    <6m       TAU      25     20     NA     NA     NA
4   No    >6m     BtheB      21     17     16     10      9
5  Yes    >6m     BtheB      26     23     NA     NA     NA
6  Yes    <6m     BtheB       7      0      0      0      0
```

The research question: Do depression levels change from baseline (pre-treatment) to 2 months follow-up post treatment? Ie, regardless of treatment group, does participating in a clinical trial have an effect on depression levels?

First six observations:

```
  bdi.pre bdi.2m
1      29      2
2      32     16
3      25     20
4      21     17
5      26     23
6       7      0
```

1. What types of variables do we have?
2. Are they "paired"?
3. How many groups are we studying?
4. How can we approach the problem?

Which scenarios represent paired measurements (or *dependent* samples)? Mark all that apply.

1. In order to study the efficacy of a new sunscreen, 50 volunteers put a standard sunscreen on their left arm and the new sunscreen on their right arm. After 3 hours, degree of redness was assessed for each arm.
2. Newer baseball stadiums are thought to attract more fans. Attendance records in the old stadium was compared to attendance records in the new stadium for 10 teams.
3. The General Social Survey (GSS) compared the number of hours worked per week among college graduates and non-college graduates.

Parameter of interest: $\mu_d = \mu_{pre} - \mu_{2m} =$ population mean difference in depression scores between baseline and two month follow-up

```
> BtheB$diff2m<-
  BtheB$bdi.pre-BtheB$bdi.2m

  bdi.pre bdi.2m diff2m
1      29      2     27
2      32     16     16
3      25     20      5
4      21     17      4
5      26     23      3
6       7      0      7
```

Three possible scenarios:

- No change:
- Improvement:
- Worsening:

Answering the research question

We can answer the research question of interest with:

- A confidence interval for μ_d
- A hypothesis test of H_0: $\mu_d = 0$
  (this is the paired t-test)
- This is essentially the same inference about a mean that we have already learned

Conditions required for *both* the CI and HT:

1. independent observations
2. normal underlying population distribution OR $n \geq 30$

Checking condition 1: independent observations

dependent measurements made on the same person

	bdi.pre	bdi.2m	diff2m
1	29	2	27
2	32	16	16
3	25	20	5
4	21	17	4
5	26	23	3
6	7	0	7

each person's change in depression score is independent from the next

The difference in depression scores between baseline and two month follow-up for person 1 is not related to the difference in depression scores between baseline and two month follow-up for person 2 (and for all other subjects in the study). This condition regarding independent observations is satisfied.

Checking condition 2

The differences in depression scores has a slight right skew; however, the sample size is $n = 97 > 30$ so this condition is satisfied.

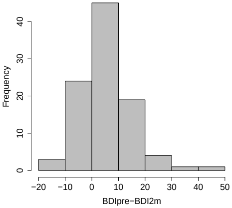
```
> favstats(BtheB$diff2m)
 min Q1 median Q3 max     mean       sd  n missing
 -17  0      5 11  41 6.237113 9.474498 97       3
```

Summary data needed for both CI and HT:

$n =$

$\bar{x}_d =$

$s_d =$



95% confidence interval for μ_d

$$\bar{x}_d \pm t^* \times \frac{s_d}{\sqrt{n}}$$

based on a 95% conf. level
and $df = n - 1 = 97 - 1 = 96$
$t^* = 1.98$ (from R)

$se = s_d / \sqrt{n} = 9.47/\sqrt{97} = 0.96$

$6.23 \pm 1.98 \times 0.96$

6.23 ± 1.90

95% CI for μ_d: (4.33,8.13)

What is the margin of error for this confidence interval?

1. 0.05
2. 0.96
3. 1.90
4. 1.98
5. 3.80

Interpretation

- Literal interpretation of the interval:
- Does it include zero?
- What does that mean?
- Answer the research question:

The 95% CI for μ_d is (4.33,8.33). Which of the following would be a *plausible* p-value when testing the hypotheses H_0: $\mu_d = 0$ vs H_a: $\mu_d \neq 0$?

1. 0.03
2. 0.23
3. 0.53
4. 0.73
5. not enough information to determine

Paired t-test

H_0: $\mu_d = 0$ vs H_a: $\mu_d \neq 0$

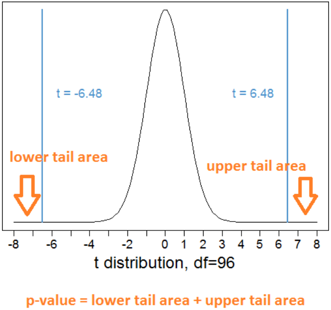
$$t = \frac{\bar{x}_d - \mu_0}{s_d/\sqrt{n}} = \frac{6.23 - 0}{9.47/\sqrt{97}} = 6.47$$

This t test statistic follows a t distribution with 96 $(n - 1)$ degrees of freedom. The p-value is found by a two-tailed area under the t distribution.

Results in R

```
#These two commands give equivalent results.  Why?

> t.test(BtheB$diff2m,mu=0)

> t.test(BtheB$bdi.pre,BtheB$bdi.2m,paired=TRUE)


        One Sample t-test

data:  BtheB$diff2m
t = 6.4836, df = 96, p-value = 3.869e-09
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 4.327579 8.146647
sample estimates:
mean of x
 6.237113
```

Interpret the p-value



If there really was no difference in average depression scores between baseline and 2 months (ie, if H_0 true), then the probability that we would observe a test statistic less than -6.48 or greater than 6.48 is really small (3.869×10^{-9}). This presents evidence against H_0.

Conclusion in context

1. **Decision about H_0:**

2. **Statement about the parameter tested in context of the research question:**

3. **Provide a deeper connection of how this relates to the research question:**

What type of error could we have committed here?

1. a Type I error
2. a Type II error
3. either
4. neither

Meaningful?

```
        One Sample t-test

data:  BtheB$diff2m
t = 6.4836, df = 96, p-value = 3.869e-09
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 4.327579 8.146647
sample estimates:
mean of x
 6.237113
```

The results are statistically signficant. But is it meaningful?

Suppose I am dealing with paired data from a randomized clinical trial regarding weight before and after a treatment intending to assist in weight loss. μ_d represents the average of post-treatment weight minus the pre-treatment weight. A 95% CI for μ_d is (0.5, 6.7).

What can we infer from this CI?

1. the treatment is not effective because the CI includes 1, indicating that there is no difference in weight loss before and after treatment
2. the treatment is effective for weight loss because the post-treatment weights are on average less than the pre-treatment weights
3. the treatment is not effective for weight loss because treatment is causing patients to gain weight, rather than lose weight, on average
4. there is not enough information to determine whether or not the treatment is effective

The Data

Paired t-test

Two sample t-test

ANOVA

Summary

The research question: Do depression levels differ at baseline (`bdi.pre`) between patients who were and were not already on antidepressant drugs (`drugs`)? OR: Is there an association between baseline depression scores and whether or not patients take antidepressant drugs?

First six observations:

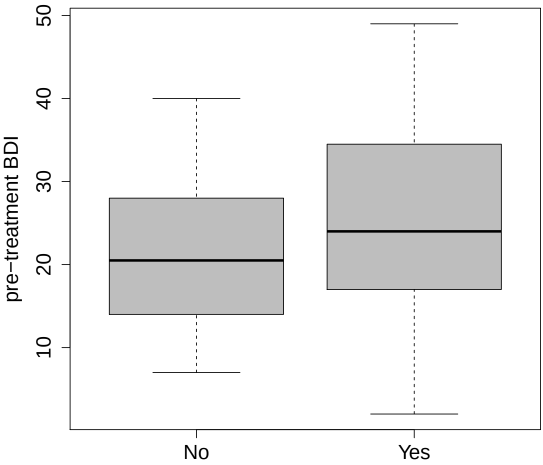
```
  drug bdi.pre
1   No      29
2  Yes      32
3  Yes      25
4   No      21
5  Yes      26
6  Yes       7
```

1. What types of variables do we have?
2. Are they "paired"?
3. How many groups are we studying?
4. How can we approach the problem?

Which figure would be appropriate to begin to visually assess if there is an association between depression scores and whether or not patients take antidepressant drugs?

1. dot plot
2. histogram
3. single boxplot
4. side by side boxplot
5. barplot

Is there an association?



Which of these scenarios represent a two sample comparison?

1. We take a random sample of 100 Cal Poly students and test if the average GPA differs from 3.3.
2. We take a random sample of 100 Cal Poly students and test if the average GPA differs between males and females.
3. We take a random sample of 100 Cal Poly students and test if the average GPA changes before and after attending a workshop on study habits.
4. We take a random sample of 100 Cal Poly students and test if the average GPA differs by political affiliation (democrat, republican, independent).

Parameters of interest:

μ_{yes} =population mean baseline BDI score among patients who are on antidepressants

μ_{no} =population mean baseline BDI score among patients who are not on antidepressants

First six observations:

```
  drug bdi.pre
1   No      29
2  Yes      32
3  Yes      25
4   No      21
5  Yes      26
6  Yes       7
```

Three possible scenarios:

- No difference:
- Difference:
- Difference:

Answering the research question

We can answer the research question of interest with:

- A confidence interval for $\mu_{yes} - \mu_{no}$
- A hypothesis test of H_0: $\mu_{yes} = \mu_{no}$
  (this is the two sample t-test)

Conditions required for *both* the CI and HT:

1. independent observations (in each of the two groups)
2. normal underlying population distribution OR
   $n \geq 30$ in *each* group

Possible outcomes

If we fail to reject H_0: $\mu_1 = \mu_2$ or if the 95% CI for $\mu_1 - \mu_2$ includes 0, then

- What is the relationship between the population mean depression scores?
- Is there evidence of an association between pre-treatment BDI score and antidepressant drug use?

If we reject H_0: $\mu_1 = \mu_2$ or if the 95% CI for $\mu_1 - \mu_2$ does not include 0, then

- What is the relationship between the population mean depression scores?
- Is there evidence of an association between pre-treatment BDI score and antidepressant drug use?

The data

```
> favstats(BtheB$bdi.pre~BtheB$drug)
  BtheB$drug min Q1 median    Q3 max     mean        sd  n missing
         No    7 14   20.5 27.50  40 21.55357  8.974549 56       0
        Yes    2 17   24.0 34.25  49 25.59091 12.577792 44       0
```

Summary data needed for both CI and HT:

<u>Yes group</u>

$n_1 =$

$\bar{x}_1 =$

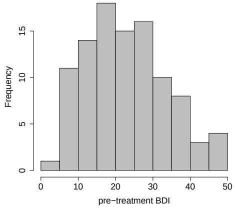
$s_1 =$

<u>No group</u>

$n_2 =$

$\bar{x}_2 =$

$s_2 =$



Which of the following data would provide the *most* evidence against H_0: $\mu_1 = \mu_2$?

1. $\bar{x}_1 = 25$, $\bar{x}_2 = 19$
2. $\bar{x}_1 = 25$, $\bar{x}_2 = 21$
3. $\bar{x}_1 = 25$, $\bar{x}_2 = 23$
4. $\bar{x}_1 = 25$, $\bar{x}_2 = 27$
5. $\bar{x}_1 = 25$, $\bar{x}_2 = 29$

95% confidence interval for $\mu_1 - \mu_2$

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \times se \text{ where } se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- The df for t^* is a complicated formula that you do not need to know. The t^* for a 95% CI with $df = 74.9$ is 1.99.
- We are calculating this CI under the assumption of *unequal variances*.

Hypothesis test of H_0: $\mu_1 = \mu_2$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{se} \text{ where } se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

We are calculating this test statistic under the assumption of *unequal variances*.

Interpretation

- Literal interpretation of the interval:
- Does it include zero?
- What does that mean?
- Answer the research question:

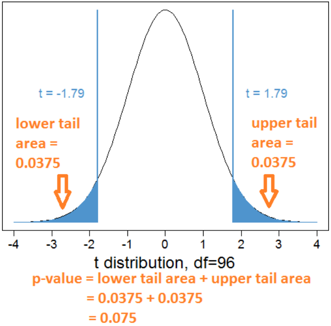
Two sample t-test assuming unequal variances

$$H_0\colon\ \mu_1 = \mu_2 \text{ vs } H_a\colon\ \mu_1 \neq \mu_2$$

```
> t.test(BtheB$bdi.pre~BtheB$drug)

        Welch Two Sample t-test

data:  BtheB$bdi.pre by BtheB$drug
t = -1.7995, df = 74.911, p-value = 0.07597
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.5069019  0.4322266
sample estimates:
 mean in group No mean in group Yes
         21.55357          25.59091
```

* Note the order of subtraction in the R output: R analyzes this as $\mu_{no} - \mu_{yes}$. You know this because the sample mean for group `No` is presented first, and the sample mean for group `Yes` is presented second.

Interpret the p-value



If average depression scores among patients who do and do not take anti-depressents really is the same (ie, if H_0 true), then the probability that we would observe a test statistic less than -1.79 or greater than 1.79 is 0.075. This probability isn't that small, and does not present evidence against H_0.

Conclusion in context

1. **Decision about H_0:**

2. **Statement about the parameter tested in context of the research question:**

3. **Provide a deeper connection of how this relates to the research question:**

Which of the following are <u>true</u>? Mark <u>all</u> that apply.

1. We could have committed a Type I error.
2. We could have committed a Type II error.
3. There is evidence of an association between antidepressant drug use and depression scores.
4. There is no evidence of an association between antidepressant drug use and depression scores.
5. We have evidence that using antidepressant drugs causes people to have higher depression scores.

A new variable

Suppose we create a variable defining patient history before they enrolled in the clinical trial. The variable is defined as:

1. `history` = 1: `drug` = no; `length` < 6m
2. `history` = 2: `drug` = no; `length` > 6m
3. `history` = 3: `drug` = yes; `length` < 6m
4. `history` = 4: `drug` = yes; `length` > 6m

	drug	length	bdi.pre	history
1	No	>6m	29	2
2	Yes	>6m	32	4
3	Yes	<6m	25	3
4	No	>6m	21	2
5	Yes	>6m	26	4
6	Yes	<6m	7	3

The `history` variable is...

1. quantitative
2. categorical

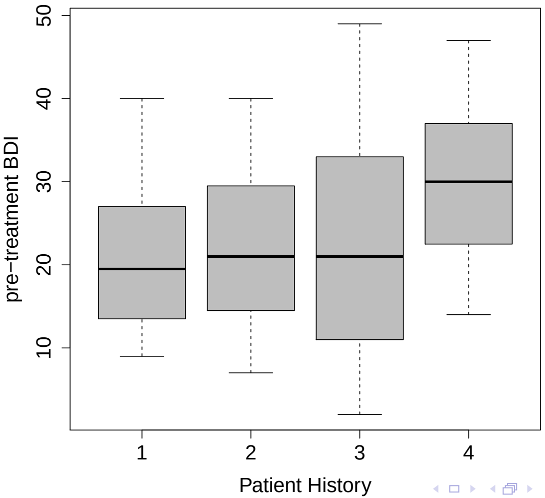
The research question: Do patients' baseline depression levels differ by patient history? OR Is there an association between baseline depression score and patient history?

First six observations:

	drug	length	bdi.pre	history
1	No	>6m	29	2
2	Yes	>6m	32	4
3	Yes	<6m	25	3
4	No	>6m	21	2
5	Yes	>6m	26	4
6	Yes	<6m	7	3

1. What types of variables do we have?
2. Are they "paired"?
3. How many groups are we studying?
4. How can we approach the problem?

Is there an association?



Motivation

- We have already learned a hypothesis test for comparing two population means (two-sample t-test)

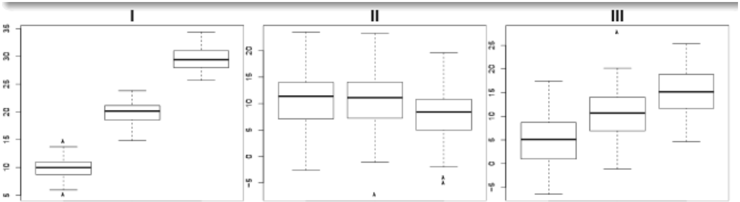
  $H_0 : \mu_1 = \mu_2$

  $H_a : \mu_1 \neq \mu_2$

- When we are interested in comparing >2 groups (and hence, >2 means) we can use ANOVA (analysis of variance)

  $H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots \mu_g$ for g groups

  H_a: at least one mean is different than the others



Which plots show groups with means that are most and least likely to be significantly different from each other?

1. most I; least II
2. most II; least III
3. most I; least III
4. most III; least II
5. most II; least I

Conclusion

$H_0: \mu_1 = \mu_2 = \mu_3 = \ldots \mu_g$ vs
H_a: at least one mean different than the others

If p-value $> \alpha$

- fail to reject H_0
- there is not sufficient evidence to suggest that population means differ significantly
- there is no evidence of an association between the two variables

If p-value $\leq \alpha$

- reject H_0
- there is evidence that at least one population means differs from the others
- there is evidence of an association between the two variables
- cannot determine which population means differ (yet)

Conditions

1. observations are independent (in each of the g groups)

   This means that the baseline depression scores are independent from patient to patient within each of the four patient history groups.

2. normal underlying population distribution OR $n \geq 30$ in each group

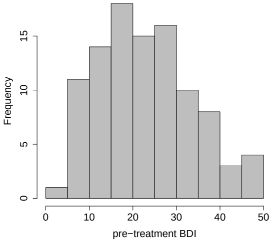
   This means that the baseline depression scores are approximately normally distributed OR $n \geq 30$ in each of the four patient history groups.

3. each group has (about) the same variability

   This means that the standard deviation of the baseline depression scores is about the same in each of the four patient history groups.

Checking conditions

```
> favstats(BtheB$bdi.pre~BtheB$history)
history min    Q1 median    Q3 max     mean        sd  n missing
      1   9 13.75   19.5 27.00  40 20.87500  8.931198 24       0
      2   7 14.75   21.0 29.25  40 22.06250  9.115522 32       0
      3   2 11.00   21.0 33.00  49 22.12000 13.185598 25       0
      4  14 22.50   30.0 37.00  47 30.15789 10.361591 19       0
```



The Mechanics

For ANOVA...

you do need to know

- how to set up H_0 and H_a
- state a conclusion regarding H_0 and H_a based on the p-value

you don't need to know

- how to calculate the test statistic
- how to shade areas for the p-value

Results in R

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ vs H_a: at least one mean differs

```
> results <- aov(BtheB$bdi.pre ~ BtheB$history)
> summary(results)
              Df Sum Sq Mean Sq F value Pr(>F)
BtheB$history  3   1118   372.8   3.404 0.0208
Residuals     96  10516   109.5
```

The p-value is 0.0208. What is your conclusion?

1. reject H_0; we have evidence that the mean BDI differs in all 4 history groups
2. reject H_0; we have evidence that at least one mean BDI differs from the others
3. fail to reject H_0; we have evidence that the mean BDI is the same in all 4 history groups
4. fail to reject H_0; we do not have evidence that the mean BDI differs in the 4 history groups

Multiple Comparisons

- When you reject $H_0 : \mu_1 = \mu_2 = \mu_3$ in one-way ANOVA, it is of interest to determine exactly which means differ and to what extent they differ
- This results in *pairwise comparisons*. If you have 4 groups, there are 6 pairwise comparisons:
  1. group 1 vs group 2
  2. group 1 vs group 3
  3. group 1 vs group 4
  4. group 2 vs group 3
  5. group 2 vs group 4
  6. group 3 vs group 4
- For each comparison, we can answer this by either testing $H_0 : \mu_1 = \mu_2$ or by calculating a confidence interval for $\mu_1 - \mu_2$

BUT...

But this means we are doing *multiple* testing, or calculating *multiple* confidence intervals

- this inflates the *overall* Type I error rate of the multiple tests taken together
- this deflates the *overall* coverage of the confidence intervals taken together

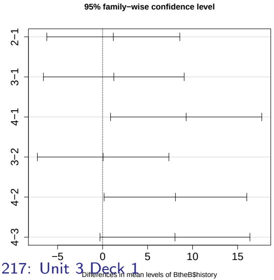
Number of tests	Chance of at least one Type I error
1	5%
3	14.3%
5	22.6%
10	40.1%
20	64.2%
50	92.3%
100	99.4%

Multiple testing techniques

- The goal of multiple testing/confidence interval techniques is to **control** the overall Type I error rate in a *set* of tests or the overall confidence level in a *set* of confidence intervals
- There are many techniques available
  - Fisher method
  - Bonferroni method
  - Tukey method
- For STAT 217, we will utilize the **Tukey method** for confidence intervals for pairwise comparisons.
  - gives overall confidence level very close to desired level
  - confidence intervals are slightly narrower than other methods
  - the formula is complex and you do not need to know it
  - you can get the results in R

Multiple Comparisons Results

```
> TukeyHSD(results)
> plot(TukeyHSD(results))
         diff         lwr       upr     p adj
2-1 1.187500 -6.2017875  8.576787 0.9749056
3-1 1.245000 -6.5750868  9.065087 0.9755714
4-1 9.282895  0.8797670 17.686022 0.0243165
3-2 0.057500 -7.2468503  7.361850 0.9999968
4-2 8.095395  0.1699716 16.020818 0.0433700
4-3 8.037895 -0.2906418 16.366431 0.0626362
```



How many pairwise comparisons indicate differences between the population means?

0 1 2 3 4 5 6

4-1 Results

```
         diff       lwr       upr     p adj
4-1 9.282895  0.8797670 17.686022 0.0243165
```

We have evidence of a difference in mean BDI score among patients in history group 4 versus 1. But is this meaningful?

Interpreting the *p*-value

The *p*-value represents the **strength** of the **evidence**:

- small *p*-values mean you have strong evidence of an association between two variables
- small *p*-values do not mean you have evidence of a strong association between two variables
- large *p*-values mean there is no evidence of an association

Other measures represent the **strength** of the **association**:

- difference of means: $(\bar{x}_1 - \bar{x}_2)$

The **strength** of the **association** can help you assess if an observed difference is meaningful.

A researcher investigated if eye color of the mother (brown vs blue) is associated with birth weight of a baby (in pounds).

```
 Two Sample t-test

data:  dat$bwt by dat$eye
t = 2.7272, df = 29998, p-value = 0.006391
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.008880031 0.054255550
sample estimates:
 mean in group blue mean in group brown
           7.543471            7.511903
```

What is the best conclusion from this analysis? We (do/do not) have strong evidence that weight of babies differs by mother's eye color, and the effect of eye color is (strong/weak).

1. do; strong
2. do not; strong
3. do; weak
4. do not; weak

Different methods

Method	**Use**	**Variables**	**Estimation**	**Testing**
Single mean *(one-sample t-test)*	quantitative response in single group	one quantitative variable	CI for μ	H_0: $\mu = \mu_0$
*Two means *(two-sample t-test)*	quantitative response in two groups	one quantitative variable and one categorical variable	CI for $\mu_1 - \mu_2$	H_0: $\mu_1 = \mu_2$
Dependent means *(paired t-test)*	quantitative response measured on same observation	two paired quantitative variables	CI for μ_d	H_0: $\mu_d = 0$
*ANOVA	quantitative response in > 2 groups	one quantitative variable and one categorical variable	Tukey pairwise intervals	H_0: $\mu_1 = \mu_2 = \cdots = \mu_g$

*The starred methods can answer the question "Is there an association?" If we reject H_0, then we conclude that some sort of association is present in the two variables.